

# SKETCH ENGINE

An introduction

*Language is never ever, ever, random* (Kilgarriff)

## CORPUS LINGUISTICS

*“CL is not a branch of linguistics in the same sense as syntax, semantics, SLX, and so on. All of these disciplines concentrate on describing/explaining some aspect of language use. CL in contrast is a methodology rather than an aspect of language requiring explanation or description. A corpus based approach can be taken to many aspects of linguistic enquiry.”*

(McEnery & Wilson, 1996: 2)

[HTTPS://WWW.SKETCHENGINE.CO.UK/](https://www.sketchengine.co.uk/)

- 400 ready-to-use corpora in 90+ languages
- <http://blogs.cardiff.ac.uk/linc/> - for links and these slides.

# SKELL



## examples, collocations and thesaurus for learners of English

### What is SkELL?

SkELL (Sketch Engine for Language Learning) is a simple tool for students and teachers of English to easily check whether or how a particular phrase or a word is used by real speakers of English.

No registration or payment required. Just type a word and click a button.

All examples, collocations and synonyms were identified automatically by ingenious algorithms and state-of-the-art software analysing large multi-billion samples of text. No manual work was involved.

### SkELL vs. Google Search

SkELL finds [good examples](#) of the word or phrase useful for language learners.

Google Search finds [web pages](#) with information about the topic specified by the word or phrase.

Try SkELL

for learners of English

ruSkELL for Russian

для изучающих русский язык

csSkELL for Czech

pro studenty češtiny

# OPEN CORPORA AND SOME RECENT ADDITIONS

## EUR-Lex Judgements Corpus

## Extended corpus of English broadsheets

## New academic English corpus

**DOAJ** DIRECTORY OF  
OPEN ACCESS  
JOURNALS

A new corpus of academic corpus was collected from the [Directory of Open Access Journals](#) words.

The DOAJ corpus contains title, country, year of publication, and is searchable to a very narrow parameter. It currently offers only the English version, but more are planned.

## New corpus from the environment domain

You are here: [Home](#)

### modifiers of "climate"

	<u>6,801</u>	26.43
global +	<u>419</u>	9.55
global climate .		
humid +	<u>107</u>	8.87
warm +	<u>174</u>	8.85
wave +	<u>888</u>	8.81
wave climate		
tropical +	<u>203</u>	8.79
wet +	<u>112</u>	8.66
wet climate		
temperate	<u>91</u>	8.60
in temperate climates		

The [LexiCon Research Group at the University of Granada](#) developed and provided their highly specialised English [EcoLexicon](#) corpus built up of environmental texts. The corpus is hosted as an open corpus and is [freely accessible](#) even without a Sketch Engine account.


The corpus is a great source for searching keywords and terms from the field of environment. The EcoLexicon enables the user to search in a specific language variant (British, American, etc.), sort results by a country or year of publication, even specify a domain or genre.

[EcoLexicon corpus in detail](#)

See the most typical [collocations](#) for the noun *climate*.

[show collocations](#)

# GETTING HELP


Sketch Engine


HomeNews & EventsGet Sketch EngineUser GuideFAQAbout usContact


Search icon


User guide


You are here: Home / User guide


 **Quick start guide**  
main features in 1 minute


 **User manual**  
basic and advanced


 **Documentation**  
for expert users

 **How do I ...?**  
user guide arranged by task

 **FAQ**  
user space, subscription and payment

 **Glossary**  
terminology in one place

 **Video**  
self-study video resources

 **Attend training**  
face-to-face

## A guide to Sketch Engine for...

Sketch Engine is for anyone working with or interested in or learning a language. Learn which features are the best for you.

Translators

Teachers

Terminologists

Students

Lexicographers

Historians

## KEY FEATURES OF SKETCH ENGINE

- Create your own corpus
- Word sketch
- Word Sketch Differences & Bilingual Word Differences
- Automatic term extraction
- Parallel Corpora
- ...

## BASIC SEARCHES

- Search for a word or phrase
- Using EcoLexicon English (Environment)
- Go to <https://the.sketchengine.co.uk/open/>



# KWIC CONCORDANCE LINES

Sketch文Engine

⇅ 🔍

📄 EcoLexicon English (Environment)

Home

Search

Word list

Word sketch

Thesaurus

Sketch diff

Corpus info

My jobs

User guide ↗

Simple query:

Make Concordance

[Query types](#)

[Context](#)

[Text types](#)

?

# WORD LIST

- What do you think is the most frequently used word in this corpus?
- Let's check...

The screenshot shows the 'Word list options' page in the Sketch Engine interface. The left sidebar contains navigation links: Home, Search, Word list (highlighted), Word sketch, Thesaurus, Sketch diff, Corpus info, My jobs, and User guide. Below these are 'All words', 'All lemmas', and 'Find x'. The main content area is titled 'Word list options' and includes the following sections:

- Subcorpus:** A dropdown menu set to 'None (whole corpus)' with links for 'info' and 'create new'.
- Search attribute:** A dropdown menu set to 'word'.
- Filters:** Two checkboxes: 'use n-grams. Value of n: from 2 to 2' and 'hide/nest sub-n-grams'.
- Filter options:**
  - Filter word list by:** A text input field for 'Regular expression'.
  - Minimum frequency:** A text input field set to '5'.
  - Maximum frequency:** A text input field set to '0' with a note '(0 = no maximum frequency)'.
  - Whitelist:** A 'Choose file' button, 'No file chosen' text, and a 'Clear' button.
  - Blacklist:** A 'Choose file' button, 'No file chosen' text, a 'Clear' button, and a 'format' link.
  - ☐ Include non-words
- Output options:**
  - Frequency figures:** Radio buttons for 'Hit counts' (selected), 'Document counts', and 'ARF'.
  - Output type:** Radio buttons for 'Simple' (selected) and 'Keywords'.
  - Reference (sub)corpus:** A dropdown menu set to 'English Web 2013 (enTenTen13)' and another dropdown set to '(whole corpus)'.
  - Prefer:** A slider between 'rare words' and 'common words', currently positioned towards 'rare words'.
  - Change output attribute(s):** Three empty dropdown menus.
  - A note: 'You can select one or more output attributes. Please note that this option can be time-consuming.'
- Make word list** button at the bottom.

# COLLOCATION

- What other words co-occur with WATER? WIND? AIR?

Sketch Engine water EcoLexicon English (Environment)

Home  
Search  
Word list  
Word sketch  
Thesaurus  
Sketch diff  
Corpus info  
My jobs  
User guide

Save  
Make subcorpus  
View options  
KWIC  
Sentence  
Sort  
Left  
Right  
Node  
Shuffle  
Sample  
Filter  
Sub-hits  
1st hit in doc  
Frequency  
Node tags  
Node forms  
Text types  
Collocations  
Visualize

Query **water** 90,533 (3,163.70 per million)

Page 1 of 4,527 Go [Next](#) | [Last](#)

**doc#0** ( rubble mound and gravity types ) increase with **water** depth , requiring a great amount of  
**doc#0** . In addition , they prevent the circulation of **water** and so deteriorate the water quality near the  
**doc#0** the circulation of water and so deteriorate the **water** quality near the coast. In some places , they  
**doc#0** upper part is impermeable and extends above the **water** level to some distance below sea level. The  
**doc#0** flow exchange between the partially enclosed **water** body and the open sea. The present  
**doc#0** barrier near the free surface extends above the **water** level to some distance below sea level and  
**doc#0** facilities ( wave flume dimensions and used **water** depths ) , wave type and main investigated wave  
**doc#0** sketched in Fig. 1 , in which h is the constant **water** depth in still water ; D [ m ] is the draft of the  
**doc#0** 1 , in which h is the constant water depth in still **water** ; D [ m ] is the draft of the upper part ; b [ m ] is the  
**doc#0** measured vertically upwards from the still **water** level. A regular wave train with wave height H<sub>i</sub> [  
**doc#0** study. The proposed breakwater can be used for **water** depths ranging from 10 m to 15 m. However , the  
**doc#0** is carried out in the laboratory for a constant **water** depth of 0.50 m , slots width and gap of 0.02 m , and  
**doc#0** from 0.9 to 1.9 s. These ranges correspond to 15 m **water** depth , 0.6 m slots width and gap , and 4.9 to 10.4 s  
**doc#0** experiments were carried out with a constant **water** depth ( h ) of 0.5 m and with generator motions  
**doc#0** is 6 m each. 4.4. Wave height measurements The **water** level variations which resulted from the  
**doc#0** D/h. This can be explained by considering the **water** particle motions. As h/L increases the water  
**doc#0** water particle motions. As h/L increases the **water** particle velocity and acceleration increases  
**doc#0** the wave comes across the breakwater model , the **water** particle velocity and acceleration suddenly  
**doc#0** ). In addition , as D/h increases , the area which **water** path through decreases then the transmitted  
**doc#0** wave motion is minimal in the lower part of the **water** column for short waves ( Huang 2007 ). 5.5. Model

Page 1 of 4,527 Go [Next](#) | [Last](#)

Sketch Engine water EcoLexicon English (Environment)

Home  
Search  
Word list  
Word sketch  
Thesaurus  
Sketch diff  
Corpus info  
My jobs  
User guide

concordance  
Frequency  
Node tags  
Node forms  
Text types

Collocation candidates

Attribute: word In the range from: -5 to: 5

Minimum frequency in corpus: 5

Minimum frequency in given range: 3

T-score	T-score
MI	MI
MI3	MI3
log likelihood	log likelihood
min. sensitivity	min. sensitivity

Show functions: logDice Sort by: logDice

[Make candidate list](#) [Save options](#)



## MORE ADVANCED SEARCHES


### WORD – LEMMA - TAG

TASK	CQL CODE	RESULT
find examples of “went”	[word="went"]	concordance of the word went
find examples of all forms of go	[lemma="go"]	concordance of go, goes, going, gone, went
find examples of all words tagged with the tag NP	[tag="NP"]	concordance of various words tagged as NP


# MORE ADVANCED SEARCHES

## WORD – LEMMA - TAG


Sketch Engine    EcoLexicon English (Environment)

Home  
Search  
Word list  
Word sketch  
Thesaurus  
Sketch diff  
Corpus info  
My jobs  
User guide 


Simple query:

[Query types](#) [Context](#) [Text types](#) 


Query type ☐ simple ☐ lemma ☐ phrase ☐ word ☐ character ☒ CQL

Lemma:  PoS: unspecified 

Phrase:

Word form:  PoS: unspecified  ☐ match case

Character:

CQL:  Default attribute: word 

[Tagset summary](#) [CQL builder](#)

# WORD SKETCHES

## test/experiment

(noun) Alternative PoS: [verb](#) (freq: 1,427)

British Academic Written English Corpus (BAWE) freqs = [2,458](#) | [2,186](#)

test 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 experiment

and/or	240	181	0.10	0.08
test	<a href="#">20</a>	0	10.4	--
score	<a href="#">7</a>	0	9.7	--
ultrasound	<a href="#">4</a>	0	9.0	--
x-ray	<a href="#">3</a>	0	8.5	--
examination	<a href="#">3</a>	0	8.2	--
analysis	<a href="#">6</a>	0	7.9	--
result	<a href="#">5</a>	0	7.6	--
data	<a href="#">3</a>	0	7.2	--
sample	<a href="#">4</a>	<a href="#">3</a>	8.0	7.8
time	0	<a href="#">3</a>	--	6.0
value	0	<a href="#">3</a>	--	6.4
theory	0	<a href="#">6</a>	--	7.6
observation	0	<a href="#">3</a>	--	8.3
concentration	0	<a href="#">3</a>	--	8.3
participant	0	<a href="#">4</a>	--	8.8
experiment	0	<a href="#">14</a>	--	10.3

subject_of	242	307	0.10	0.14
confirm	<a href="#">7</a>	0	8.9	--
give	<a href="#">9</a>	<a href="#">4</a>	7.3	6.1
indicate	<a href="#">11</a>	<a href="#">5</a>	8.3	7.1
consist	<a href="#">6</a>	<a href="#">3</a>	8.0	6.8
measure	<a href="#">6</a>	<a href="#">5</a>	8.7	8.2
produce	<a href="#">3</a>	<a href="#">4</a>	6.3	6.6
carry	<a href="#">5</a>	<a href="#">7</a>	7.9	8.3
involve	<a href="#">14</a>	<a href="#">19</a>	8.6	9.0
show	<a href="#">27</a>	<a href="#">38</a>	8.0	8.5
require	<a href="#">3</a>	<a href="#">5</a>	6.4	7.0
suggest	<a href="#">3</a>	<a href="#">7</a>	5.6	6.8
use	<a href="#">9</a>	<a href="#">23</a>	6.4	7.7
make	0	<a href="#">4</a>	--	5.4
find	0	<a href="#">3</a>	--	6.5
include	0	<a href="#">6</a>	--	7.0
present	0	<a href="#">3</a>	--	7.1
highlight	0	<a href="#">3</a>	--	7.1
support	0	<a href="#">5</a>	--	7.2
run	0	<a href="#">3</a>	--	7.3
take	0	<a href="#">11</a>	--	7.4
look	0	<a href="#">5</a>	--	7.6
demonstrate	0	<a href="#">5</a>	--	7.7
aim	0	<a href="#">7</a>	--	8.3
prove	0	<a href="#">12</a>	--	9.0
investigate	0	<a href="#">9</a>	--	9.4

adj_subject_of	55	27	0.02	0.01
useful	<a href="#">4</a>	0	8.8	--
accurate	0	<a href="#">3</a>	--	9.4

object_of	492	384	0.20	0.18
pass	<a href="#">13</a>	0	8.9	--
reset	<a href="#">6</a>	0	8.6	--
apply	<a href="#">15</a>	0	8.4	--
stretch	<a href="#">5</a>	0	8.2	--
stand	<a href="#">5</a>	0	8.2	--
devise	<a href="#">5</a>	0	8.1	--
execute	<a href="#">5</a>	0	8.0	--
fail	<a href="#">4</a>	0	7.8	--

subject_of	242	307	0.10	0.14
confirm	<a href="#">7</a>	0	8.9	--
give	<a href="#">9</a>	<a href="#">4</a>	7.3	6.1
indicate	<a href="#">11</a>	<a href="#">5</a>	8.3	7.1
consist	<a href="#">6</a>	<a href="#">3</a>	8.0	6.8
measure	<a href="#">6</a>	<a href="#">5</a>	8.7	8.2
produce	<a href="#">3</a>	<a href="#">4</a>	6.3	6.6
carry	<a href="#">5</a>	<a href="#">7</a>	7.9	8.3
involve	<a href="#">14</a>	<a href="#">19</a>	8.6	9.0
show	<a href="#">27</a>	<a href="#">38</a>	8.0	8.5
require	<a href="#">3</a>	<a href="#">5</a>	6.4	7.0
suggest	<a href="#">3</a>	<a href="#">7</a>	5.6	6.8
use	<a href="#">9</a>	<a href="#">23</a>	6.4	7.7
make	0	<a href="#">4</a>	--	5.4
find	0	<a href="#">3</a>	--	6.5
include	0	<a href="#">6</a>	--	7.0
present	0	<a href="#">3</a>	--	7.1
highlight	0	<a href="#">3</a>	--	7.1
support	0	<a href="#">5</a>	--	7.2
run	0	<a href="#">3</a>	--	7.3
take	0	<a href="#">11</a>	--	7.4
look	0	<a href="#">5</a>	--	7.6
demonstrate	0	<a href="#">5</a>	--	7.7
aim	0	<a href="#">7</a>	--	8.3
prove	0	<a href="#">12</a>	--	9.0
investigate	0	<a href="#">9</a>	--	9.4

## team

(noun) Alternative PoS: [verb](#) (478)

British National Corpus (BNC) freq = [22,482](#) (200.21 per million)

modifiers of "team"	13,919	0.62
management	<a href="#">433</a>	9.31
+		
management team		
football	<a href="#">207</a>	8.63
+		
football team		
project	<a href="#">166</a>	8.35
+		
the project team		
england	<a href="#">143</a>	8.05
+		
the england team		
research	<a href="#">164</a>	7.83
+		
the research team		
rescue	<a href="#">98</a>	7.76
+		
mountain rescue team		
display	<a href="#">91</a>	7.60
+		
the national display team		
cup	<a href="#">96</a>	7.45
+		
cup team		
design	<a href="#">87</a>	7.38
+		
the design team		
nouns and verbs modified by "team"	3,166	0.14
spirit	<a href="#">112</a>	9.15
+		
team spirit		
mate	<a href="#">53</a>	8.75
+		
his team mates		
leader	<a href="#">133</a>	8.26
+		
team leader		
coach	<a href="#">40</a>	8.09
+		
the team coach		
manager	<a href="#">133</a>	8.05
+		
team manager		
member	<a href="#">197</a>	8.01
+		
team members		
effort	<a href="#">72</a>	7.94
+		
a team effort		
championship	<a href="#">49</a>	7.77
+		
team championship		
selection	<a href="#">38</a>	7.73
+		
the selection team		
verbs with "team" as object	4,616	0.21
lead	<a href="#">205</a>	8.48
+		
head	<a href="#">63</a>	8.26
+		
team headed by		
join	<a href="#">113</a>	8.04
+		
pick	<a href="#">47</a>	7.79
+		
field	<a href="#">26</a>	7.43
+		
assemble	<a href="#">25</a>	7.17
+		
beat	<a href="#">34</a>	7.01
+		
negotiate	<a href="#">26</a>	7.00
+		
captain	<a href="#">18</a>	6.92
+		
send	<a href="#">55</a>	6.86
+		
strengthen	<a href="#">22</a>	6.79
+		
investigate	<a href="#">27</a>	6.77
+		
the investigating team		
select	<a href="#">27</a>	6.74
+		
visit	<a href="#">36</a>	6.53
+		
verbs with "team" as subject	6,300	0.28
win	<a href="#">98</a>	7.97
+		
team won		
play	<a href="#">105</a>	7.86
+		
work	<a href="#">109</a>	7.53
+		
team working		
lose	<a href="#">40</a>	6.78
+		
team lost		
consist	<a href="#">31</a>	6.78
+		
team consists of		
perform	<a href="#">27</a>	6.74
+		
compete	<a href="#">22</a>	6.70
+		
teams competing in		
find	<a href="#">57</a>	6.55
+		
team found		
comprise	<a href="#">21</a>	6.46
+		
team comprising		
prepare	<a href="#">22</a>	6.45
+		
105	<a href="#">6.36</a>	
"team" and/or...	2,244	0.10
football	<a href="#">12</a>	7.15
+		
cast	<a href="#">8</a>	6.75
+		
search	<a href="#">9</a>	6.71
+		
group	<a href="#">31</a>	6.55
+		
squad	<a href="#">7</a>	6.55
+		
individual	<a href="#">12</a>	6.41
+		
husband	<a href="#">12</a>	6.37
+		
husband and wife team		
player	<a href="#">10</a>	6.35
+		
supporter	<a href="#">7</a>	6.19
+		
afternoon	<a href="#">7</a>	6.17
+		
fan	<a href="#">6</a>	6.11
+		
panel	<a href="#">6</a>	6.11
+		
specialist	<a href="#">6</a>	6.08
+		
sale	<a href="#">10</a>	6.07
+		
member	<a href="#">16</a>	6.01
+		
department	<a href="#">10</a>	5.93
+		
management	<a href="#">12</a>	5.91
+		
manager	<a href="#">13</a>	5.88
+		

## house

(noun)

ukWaC freq = [391,778](#) (251.18 per million)

## Haus

(noun)

deTenTen [2013] freq = [7,264,685](#) (364.72 per million)

Use another candidate translation: [erinnern](#) [Hausarrest](#) [Ordnung](#) [Plenum](#) [hoch](#) [daran](#) [Parlament](#) [kehren](#) [mitteilen](#)  
Click on collocates to access reciprocal bilingual search

object of	96,897	2.10
terrace	<a href="#">1,667</a>	9.04
detach	<a href="#">1,737</a>	8.94
build	<a href="#">8,502</a>	8.59
buy	<a href="#">3,998</a>	8.10
board	<a href="#">853</a>	7.94
rent	<a href="#">935</a>	7.90
sell	<a href="#">2,452</a>	7.58
demolish	<a href="#">650</a>	7.54
situate	<a href="#">1,061</a>	7.39
own	<a href="#">1,284</a>	7.25
occupy	<a href="#">798</a>	7.12
move	<a href="#">2,456</a>	7.00
VerbY+SubstXDat (obj dat of)	1,476,115	3.60
wohnen	<a href="#">31,400</a>	7.14
fühlen	<a href="#">26,177</a>	6.04
fahren	<a href="#">46,582</a>	5.99
kehren	<a href="#">7,991</a>	5.64
befinden	<a href="#">30,253</a>	5.35
schicken	<a href="#">9,019</a>	5.34
sitzen	<a href="#">15,913</a>	5.25
leben	<a href="#">23,884</a>	5.21
kommen	<a href="#">144,447</a>	5.04
holen	<a href="#">10,220</a>	4.99
eilen	<a href="#">1,891</a>	4.84
rennen	<a href="#">2,939</a>	4.82
subject of	58,690	1.90
overlook	<a href="#">244</a>	6.20
stand	<a href="#">734</a>	6.12
belong	<a href="#">306</a>	6.09
rebuild	<a href="#">135</a>	5.61
date	<a href="#">266</a>	5.52
front	<a href="#">86</a>	5.39
burn	<a href="#">113</a>	4.98
line	<a href="#">84</a>	4.93
occupy	<a href="#">133</a>	4.87
collapse	<a href="#">63</a>	4.71
boast	<a href="#">75</a>	4.68
survive	<a href="#">107</a>	4.56
SubstXNom+VerbY (subj of)	322,711	0.80
beherbergen	<a href="#">1,910</a>	5.78
verfügen	<a href="#">16,744</a>	5.49
brennen	<a href="#">3,018</a>	5.48
abbrennen	<a href="#">364</a>	4.83
schmiegen	<a href="#">384</a>	4.60
erstrahlen	<a href="#">562</a>	4.56
befinden	<a href="#">15,624</a>	4.47
bestechen	<a href="#">933</a>	4.33
säumen	<a href="#">294</a>	4.17
gruppieren	<a href="#">255</a>	4.14
liegen	<a href="#">32,608</a>	3.99
einstürzen	<a href="#">187</a>	3.97

adj subject of	7,821	2.10
modifier	<a href="#">174,914</a>	1.30
modifier	<a href="#">6,611,554</a>	1.60
modifies	<a href="#">54,780</a>	0.40

# WORD SKETCH

- Create a word sketch for AIR, WIND, WATER (or any word you like)

Sketch Engine

Q

EcoLexicon English (Environment)

Home

Search

Word list

Word sketch

Thesaurus

Sketch diff

Corpus info

My jobs

User guide

Word sketch

Lemma: air

Part of speech: auto

Advanced options

Show word sketch

# SKETCH DIFFERENCE

- SIMPLE vs COMPLEX

Sketch Engine   EcoLexicon English (Environment)

Home  
Search  
Word list  
Word sketch  
Thesaurus  
**Sketch diff**  
Corpus info  
My jobs  
User guide [↗](#)

### Word sketch differences ?

Lemma:

Part of speech:

Sketch diff by: ☒ lemma

Second lemma:

☐ subcorpus

First subcorpus:  [info](#) [create new](#) ?

Second subcorpus:  [info](#) [create new](#) ?

☐ word form

First word form:

Second word form:

[Advanced options](#)



## STATISTICAL MEASURES

- some understanding of the measures used is needed

## MI SCORE

**MI score:** a measure of how strongly two words seem to associate in a corpus, based on the independent relative frequency of two words.

- 1) not dependent on the size of the corpus
  - 2) can be compared across corpora, even if the corpora are of different sizes
  - 3) gives information about its lexical behaviour, but particularly about the more idiomatic co-occurrences
  - 4) the highest MI scores tend to be less frequent words with restricted collocation.
- The strength of the collocation is **not always a reliable indication of meaningful association.**

## T-SCORE

**t-score** : a measure of how certain we can be that the collocation is the result of more than the vagaries of a particular corpus

- 1) Corpus size is important.
  - 2) cannot be compared across corpora
  - 3) gives information about the grammatical behaviour of a word
  - 4) the highest t-scores tend to be frequently used words ( whether or not they are grammatical words) that collocate with a variety of other words.
- In some instances they may require a wider span than is commonly used with respect to 'clause collocation'

# INTERPRETING COLLOCATES

## Collocation candidates for PUSS

	<u>Freq</u>	<u>T-score</u>	<u>MI</u>
<a href="#">p/n</a> puss	10	3.162	18.882
<a href="#">p/n</a> glamour	4	2.000	12.646
<a href="#">p/n</a> sour	3	1.732	12.016
<a href="#">p/n</a> Taking	3	1.731	11.589
<a href="#">p/n</a> Hello	4	1.999	10.861
<a href="#">p/n</a> Little	3	1.730	9.577
<a href="#">p/n</a> Here	5	2.232	9.253
<a href="#">p/n</a> November	3	1.723	7.646
<a href="#">p/n</a> black	3	1.723	7.542
<a href="#">p/n</a> bit	3	1.721	7.238
<a href="#">p/n</a> Britain	3	1.720	7.209

"sour" f=4109 ; "puss" f=254

"sour" only co-occur 3 times, this gives this particular collocation a very high MI score: i.e. these two words will be very strongly associated.

However, the t-score says "maybe, but we haven't seen enough evidence to be sure that the MI is right!".

The t-score is relatively low: 1.73

# INTERPRETING COLLOCATES

	Freq.	T-score	MI
<a href="#">p/n</a> pheasants	4	1.995	8.672
<a href="#">p/n</a> bottomless	9	2.993	8.670
<a href="#">p/n</a> stretchered	3	1.728	8.646
<a href="#">p/n</a> 17-0	3	1.728	8.639
<a href="#">p/n</a> torrents	5	2.230	8.636
<a href="#">p/n</a> pro-business	3	1.728	8.593
<a href="#">p/n</a> farmhouses	5	2.230	8.590
<a href="#">p/n</a> 100ft	4	1.995	8.586
<a href="#">p/n</a> madly	13	3.596	8.586
<a href="#">p/n</a> dappled	4	1.995	8.564
<a href="#">p/n</a> steadily	91	9.514	8.556
<a href="#">p/n</a> bemoan	3	1.727	8.472
<a href="#">p/n</a> seams	15	3.862	8.471
<a href="#">p/n</a> rain	329	18.087	8.468
<a href="#">p/n</a> Graff	3	1.727	8.465
<a href="#">p/n</a> Recovery	18	4.231	8.462
<a href="#">p/n</a> Moslems	4	1.994	8.453
<a href="#">p/n</a> 37.5	3	1.727	8.452
<a href="#">p/n</a> weightless	3	1.727	8.452
<a href="#">p/n</a> storeys	8	2.820	8.446
<a href="#">p/n</a> underperformed	3	1.727	8.445
<a href="#">p/n</a> gashed	3	1.727	8.438
<a href="#">p/n</a> fracturing	4	1.994	8.418
<a href="#">p/n</a> prices	827	28.673	8.415

## FALLING PRICES

f("falling") = 23,209

f("prices") = 66,352

The MI figure is not particularly high (8.415) because there is plenty of evidence of "falling" occurring without "prices" and vice versa.

Statistically the strength of association between "falling" and "prices" is much less than it was for "sour" and "puss". The t-score however is quite high at 28.673 shows it has taken into account the actual number of observations.

## A SAFE GUIDE:

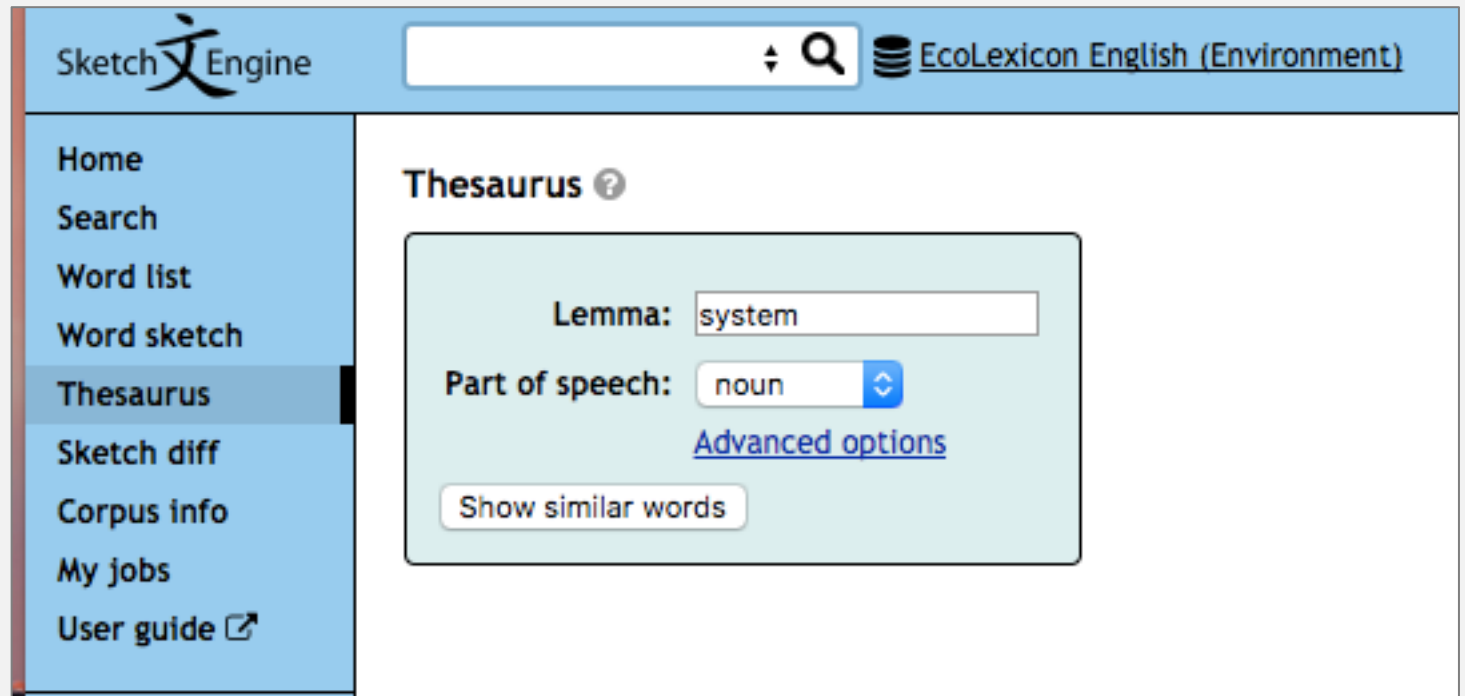
- A high T-score says: it is safe to claim that there is some non-random association between these two words.
- MI will highlight the technical terms, oddities, totally fixed phrases, etc.
- If a collocate appears in the top of both MI and T-score lists it is clearly a solid collocate

## STATISTICAL MEASURES IN SKETCH ENGINE

- MI, T-Score but also Sketch Engine's own LogDice
- LOGDICE:
  - a statistic measure based only on a frequency of words  $w_1$  and  $w_2$  and the bigram  $w_1w_2$ , it is not affected by a size of the corpus
- See <https://www.sketchengine.co.uk/documentation/statistics-used-in-sketch-engine/#logdice> for more detail on various other calculations.

# EXPLORING SOME OTHER FEATURES

- Filters
- Text Types
- Thesaurus
- Etc.



The screenshot shows the Sketch Engine Thesaurus interface. The top header bar is blue and contains the 'Sketch Engine' logo, a search bar with a magnifying glass icon, and a database selector set to 'EcoLexicon English (Environment)'. A left-hand navigation menu is also blue and lists several options: 'Home', 'Search', 'Word list', 'Word sketch', 'Thesaurus' (which is highlighted with a dark blue bar), 'Sketch diff', 'Corpus info', 'My jobs', and 'User guide' with an external link icon. The main content area is white and titled 'Thesaurus' with a help icon. It features a light blue box containing input fields for 'Lemma:' (with the text 'system') and 'Part of speech:' (with a dropdown menu showing 'noun'). Below these fields is a link for 'Advanced options' and a button labeled 'Show similar words'.



## HOW MIGHT YOU WANT TO USE SKETCH ENGINE?

- Some free time to explore ways in which you might want to use Sketch Engine

## REFERENCES

- Adam Kilgarriff: <https://www.kilgarriff.co.uk/>
  - Rich resource of papers and presentations, e.g. How Many Words are There?, and many more
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press
- Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel (2014): The Sketch Engine: ten years on.. *Lexicography* 1(1): 7–36.
- McEnery, T. & Wilson, A. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Rychlý, P. (2008). A lexicographer-friendly association score. In Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN, pp. 6–9.
- Thomas, J. 2016. *Discovering English with the Sketch Engine*. 2<sup>nd</sup> edition. Versatile.